Structural equation modeling in epidemiology

Modelagem de equações estruturais em epidemiologia

Leila Denise Alves Ferreira Amorim [1]
Rosemeire L. Fiaccone [1]
Carlos Antônio S. T. Santos [2]
Tereza Nadya dos Santos [1]
Lia Terezinha L. P. de Moraes [1]
Nelson F. Oliveira [2]
Silvano O. Barbosa [1]
Darci Neves dos Santos [3]
Letícia Marques dos Santos [3]
Sheila M. A. Matos [3]
Maurício L. Barreto [3]

[1] *Instituto de Matemática, Universidade Federal da Bahia, Salvador, Bahia.*
[2] *Departamento de Ciências Exatas, Universidade Estadual de Feira de Santana, Feira de Santana, Brasil.*
[3] *Instituto de Saúde Coletiva, Universidade Federal da Bahia, Salvador, Brasil.*

Correspondence
L. D. A. F. Amorim
Instituto de Matemática, Universidade Federal da Bahia.
Av. Adhemar de Barros s/n, Salvador, BA 40170-115, Brasil.
leiladen@ufba.br

Abstract

Structural equation modeling (SEM) is an important statistical tool for evaluating complex relations in several research areas. In epidemiology, the use and discussion of SEM have been limited thus far. This article presents basic principles and concepts in SEM, including an application using epidemiological data analysis from a study on the determinants of cognitive development in young children, considering constructs related to organization of the child's home environment, parenting style, and the child's health status. The relations between the constructs and cognitive development were measured. The results showed a positive association between psychosocial stimulus at home and cognitive development in young children. The article presents the contributions by SEM to epidemiology, highlighting the need for an a priori theoretical model for improving the study of epidemiological questions from a new perspective.

Mathematical Models; Statistical Factor Analysis; Causality

Introduction

Recent years have witnessed growing interest in structural equation modeling (SEM) in various fields of knowledge. The popularity of SEM stems mainly from the fact that researchers have realized the need to understand the complex interrelations between the multiple variables under study. Traditional statistical methods only apply to a limited number of variables, and may thus fail to deal with the sophisticated emerging theories. Another reason relates to the increased recognition of the validity and reliability of scores observed using measurement instruments, when SEM allows the explicit inclusion of measurement errors. In addition, the availability of various user-friendly computer programs allows the implementation of a wide range of models, from the simplest to the most sophisticated, providing a further reason for the widespread current use of these methodologies.

SEM is a generic analytical tool that has proven flexible and powerful for estimating parameters in an extensive family of linear models [1,2]. One important aspect of SEM is its extension for estimating measurement errors by using factors or latent variables. These models consider the inclusion of variables that are not measured directly, but through their effects, called indicators, or their observable causes. Non-measurable variables are known as latent variables or construct. SEM allows assessing how sets of observed vari-

ables define constructs and how these constructs relate to each other [3]. Such modeling also allows evaluating complex mediating mechanisms by disaggregating the effects [4,5].

Although structural equation modeling has traditionally been used more in research in the social and human sciences (particularly psychology and econometrics), in recent years a wider range of applications to health has appeared in the scientific literature. Examples include the use of confirmatory factor analysis to evaluate and compare competing models for metabolic syndrome [6] or the use of SEM to develop a causal model for the relations between symptoms and quality of life in cancer patients receiving end-of-life care [7]. Comparison of the results obtained through the use of SEM and the multiple linear regression model was used to evaluate the impact of breastfeeding on cognitive function in children [8]. More recently, SEM was used in a study to examine the relations between blood pressure in adolescence, fetal development, and maternal characteristics [9]. A literature search in PubMed using the key words *structural equation modeling, confirmatory factor analysis, structural equation*, and *path analysis* in six leading periodicals in the field of Epidemiology (*Am J Epidemiol*, *Int J Epidemiol*, *Eur J Epidemiol*, *Ann Epidemiol*, *Lancet*, and *Epidemiology*) showed that 24 articles used SEM from 2001 to 2008, and that 62.5% of these had been published since 2006.

The current study thus aimed to summarize and illustrate fundamental concepts related to structural equation modeling, considering the most commonly described methods in the literature [1,2,10,11,12,13]. The study thus includes information on terminology, the model's mathematical specifications, estimation procedures, and criteria for assessing goodness of fit. The application of SEM to epidemiological data is presented through a study associated with child's health, emphasizing the interpretation of its results.

Material and methods

Structural equation modeling

SEM includes multivariate data analysis techniques that combine aspects from multiple regression and factor analysis to simultaneously estimate a series of relations of dependency that allow defining procedures aimed at directly incorporating measurement errors into the model [5]. The variables used in SEM can be observed variables or constructed variables (not observed), called constructs or latent variables. This is one of the most important differences between SEM and other modeling techniques, since the classical data analysis procedures only model observable measurements [14]. Another interesting characteristic of the methodology is that a variable can be both a response in an equation and appear as an explanatory variable in another equation. It is also possible to specify a reciprocal effect, in which two variables affect each other through a feedback loop.

The application of SEM is based on the theory used by the researcher and requires a theoretical model specified *a priori* for explaining the multiple relations of dependency or causal relations among a set of variables. A theoretical model consists of a systematic set of relations that provide consistent and comprehensive explanations for the phenomena [12]. The relations are defined by a series of equations that describe the hypothetical structures, allowing the modeling of different types of correlations among observations, regardless of their origin (for example, causal relations, multiple responses, repeated measurements, and longitudinal designs) [15]. The theoretical model can be expressed through both equations and graphs, called path diagrams, which summarize the set of hypotheses.

Application of SEM consists of various stages, including the development of a theoretical-conceptual model, specification of the mathematical model, determination of the model's identifiability, the model's fit and evaluation of its goodness of fit [16]. Several of these procedures are described below.

• Types of variables and path diagrams

Variables in SEM can be classified in relation to various aspects in the model. As for their measurability, they can be classified as latent, measurement, and indicator variables. Variables that are not directly measurable are called latent variables or constructs and refer to theoretical concepts that cannot be observed directly. In SEM, the constructs allow the formation of relations of dependency or causal relations to be estimated by the models, and they are measured approximately by a set of observed variables. Latent variables can also be related to measurement variables in a relationship of dependency [12]. The observed variables that are used to compose a latent variable are called indicators. The researcher must justify the theoretical basis for the indicator variables because SEM only examines its empirical characteristics. For each construct that appears in the model, it is necessary to determine in advance which indicator variables are related to it.

As for the influence that one variable exerts on others, variables can be classified as exoge-

nous or endogenous [1,13]. Exogenous variables are not influenced or do not suffer the effect of other variables in the model, and are also called independent or predictive variables. As in the traditional regression model, these variables (qualitative or quantitative) are assumed to be measured without error. Endogenous or dependent variables are influenced by other variables present in the model. The structural errors represent the aggregate omitted causes from the endogenous variables, together with the measurement error. There will be an error associated with each endogenous variable in the model.

Since SEM generally includes complex models, many researchers find it more convenient to depict such models in diagram form. The portrayal called "path diagram" allows rapid visualization of the relations of interdependency in the theoretical model [12]. The path diagram is depicted by a set of geometric figures and arrows showing the types of variables (observed or latent) and the relations between them [1].

Other principles in the construction of a path diagram are: (i) X represents the indicator variables of the exogenous constructs and Y the indicator variables of the endogenous constructs; (ii) constructs are normally represented by circles or ovals; (iii) measurement variables are represented by rectangles or squares; (iv) indicator variables X and/or Y are associated with their respective constructs by arrows pointing from the constructs towards the indicator variables (Figure 1). SEM conventionally assumes that the indicator variables are dependent on the constructs. When two variables are not connected by an arrow, it does not necessarily mean that one does not affect the other, since the relationship can occur indirectly and is identified by more complex paths.

Relations of dependency and correlations, represented by bidirectional curves, can be demonstrated by the path diagram and illustrate the importance of SEM as a methodology to simultaneously estimate a large set of equations. Importantly, the construction of a path diagram involves two assumptions [1]. The first is that all the causal relations are shown in the diagram, and that the choice of these relations is consistent with the underlying SEM theory. According to this assumption, the objective of SEM is to model the relations with the smallest number of causal paths or correlations between the variables that can be justified theoretically. A second assumption regards the nature of the relations between the variables (latent or observed), assumed to be linear or which can be linearized by transformation.

- **Specification of the general structural equation model**

General SEM consists of measurement and structural sub-models, introduced by Jorëskog [17]. The structural sub-model can be shown as:

$$\eta = B\eta + \Gamma\xi + \zeta,$$

where η represents a vector mx1 of latent endogenous variables; ξ represents a vector kx1 of latent exogenus variables; B is a matrix mxm of coefficients relating the latent endogenous variables to each other; Γ is a matrix mxk of coefficients relating the endogenous variables to the exogenous variables; and ζ is a vector mx1 of structural disturbances. In the notation, B displays zeros on its main diagonal [8]. The latent variables are related to the observed variables through the measurement sub-model, which is defined separately for endogenous and exogenous variables by:

$$y = \Lambda_y \eta + \varepsilon \quad \text{and} \quad x = \Lambda_x \xi + \delta,$$

where Λ_y and Λ_x are matrices pxm and qxk, respectively, of factor loadings, and ε and δ are vectors px1 and qx1, respectively, of measurement errors in y and x. Each column in the Λ matrices generally contains a value which is set at 1 to establish the scale for the corresponding latent variable. Alternatively, this can be done by setting at zero the variances in the latent exogenous variables in the Φ matrix, which represents the covariance matrix of the exogenous variables [1]. The measurement model's fit corresponds to the application of confirmatory factor analysis (CFA) for definition of latent variables/constructs.

The model assumes that measurement errors ε and δ have expectation zero, each one with multivariate normal distribution, independent of each other and independent of the latent exogenous variables (ξ), latent endogenous variables (η), and disturbances (ζ). In addition, it assumes that the observations are sampled independently and that the latent exogenous variables (ξ) have a multivariate normal distribution. This latter assumption is unnecessary for exogenous variables that are measured without error. The structural disturbances (ζ) have expectation zero and multivariate normal distribution and are independent of the latent exogenous variables (ξ). Under these assumptions, the observed indicators, x and y, have a multivariate normal distribution:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N_{p+q}(0, \Sigma)$$

where Σ represents the indicators' population covariance matrix. This matrix is a function of the model's parameters Ω = (B,Γ,Λ_x,Λ_y,Ψ,Θ_δ,Θ_ε e Φ) , and can be expressed by:

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} =$$

$$= \begin{pmatrix} \Lambda_y(I-B)^{-1}(\Gamma\Phi\Gamma')[(I-B)^{-1}]'\Lambda_y'+\Theta_\varepsilon & \Lambda_y(I-B)^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'[(I-B)^{-1}]'\Lambda_y^x & \Lambda_x\Phi\Lambda'_x+\Theta_\delta \end{pmatrix},$$

where Φ is the covariance matrix *k*x*k* of the latent exogenous variables, Ψ is the covariance matrix *m*x*m* of disturbance terms, and Θ_ε and Θ_δ are covariance matrices of measurement errors ε and δ, respectively.

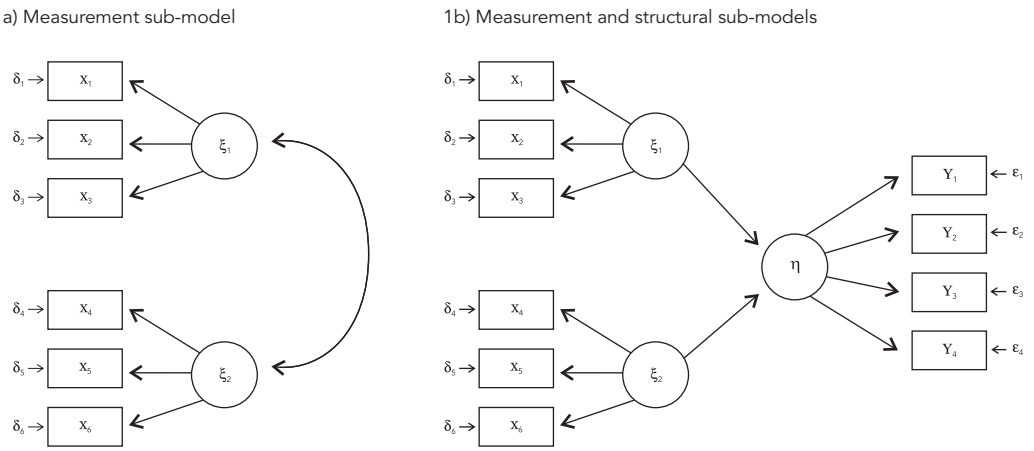
Classical path analysis can be viewed as a submodel in the more overall structure, considering $\Lambda_x = I$, $\Theta_\delta = 0$, $\Lambda_y = I$, $\Theta_\varepsilon = 0$. Likewise, to obtain matrix Σ associated with the confirmatory factor analysis, one considers B = 0, Γ = 0, Ψ = 0, $\Lambda_y = 0$ and $\Theta_\varepsilon = 0$ [5]. In any given model, restrictions will be necessary in some elements of matrix Σ. Most frequently, the restrictions include setting some of these parameters at zero. If the restrictions to the model are sufficient, estimates of maximum likelihood can be obtained for their parameters. The log-likelihood associated with this model can be defined as a function of the model's parameters, of Σ and S, the sampling covariance matrix between the observed variables.

In SEM, the procedure aims to estimate Ω so as to minimize the discrepancy function $F(S; \Sigma)$, a scalar that measures the distance between the sampling covariance matrix (S) and the adjusted covariance matrix $\hat{\Sigma}$. The two most widely used estimation methods are maximum likelihood and generalized least squares [2,13]. The likelihood logarithm can be thought of as a measurement of proximity between Σ and S. Thus, the estimates of maximum likelihood for the parameters are defined such that the two matrices are as close as possible. The asymptotic standard errors for the estimates of the parameters can be obtained from the square root of the diagonals in the information matrix. In the estimation, as stated previously, it is assumed that the structural relations between the latent exogenous and endogenous variables are linear, as are the relations between the indicator variables and the constructs associated with them.

A fundamental step in SEM is the verification models' identifiability of the latent variables, a complex problem without a simple solution. A model is said to be non-identifiable when it is not possible to find a single solution for the equation system. A necessary general condition for identifiability is the number of free parameters in the model, which cannot be greater than the number of variances and covariances between the observable variables, shown by $\frac{(p+q)x(p+q+1)}{2}$, known as the counting rule,

Figure 1

Conventions used in the path diagrams for displaying measurement and structural sub-models.



where p stands for the number of endogenous variables and q the number of exogenous variables in the model [13]. However, this condition is not sufficient, since it is easy to meet and still result in a non-identifiable model. Situations that facilitate the model's identifiability include: (i) the measurement errors are not correlated; (ii) at least two exclusive indicators exist for each latent variable; or when there is only one single indicator for a latent variable, it is measured without error; (iii) the structural model contains only observed variables [18]. The statistical software packages generally detect the attempts to adjust an under-identified model. In this case, the information matrix will be unique. Another way of verifying problems with the model's identifiability is to observe whether the estimates of the variances are very large.

A minimum number of observations are required in order for the structural equation models to be adjusted. It is recommended that the ratio between the sample size and the number of parameters to be estimated by the model be 10:1 or even 20:1, if statistical significance tests are of interest [19]. In general, SEM is less stable in small samples (60-120), thus requiring relatively large samples. The minimum sample size depends, among other things, on the model's complexity, the effect size, and the degrees of freedom [18,20].

Interpretation of the coefficients estimated by regression models is crucial for the researcher's understanding of the target relations. When the observed variables come from different arbitrary scales, it is generally necessary to use standardized coefficients to help interpret the results. The literature includes descriptions of various methods for obtaining standardized coefficients for linear regression models, including methods using measurements of variability, correlations, and standardized variables [1,21]. In the context of SEM, standardized coefficients are widely used, and are made available by all the statistical packages.

Another interesting aspect of SEM, particularly in the context of analyzing epidemiological data, is the possibility of modeling complex relations between variables, including indirect or mediated effects [22]. The simplified idea of mediation is that a third variable transmits the effect of one variable to another. Estimation of mediated effects in structural equation models and in path analysis is commonly referred to as effect decomposition and is used to explicitly identify the direct effects and the mediated effects [5]. In the health field, the intermediate variable can be called a surrogate [23].

Determination of the model's fit is complicated in this context because various goodness of fit criteria have been developed to evaluate structural equation models under different assumptions [1]. Verification of goodness of fit in SEM is not as direct as in other multivariate procedures because it is traditionally assumed that the observed variables are measured without error, and statistical tests with known distributions exist [3]. Meanwhile, the fit indices in SEM do not have a statistical significance test that identifies the correct model considering the sample data. Chi-square is the only statistical test used to evaluate the theoretical model. A statistically non-significant result for chi-square indicates that the sample's covariance matrix and the covariance matrix estimated by the model are similar. However, use of the chi-square test is known to be problematic [13,24] since it is sensitive to the sample size. It thus becomes difficult to reject the null hypothesis in studies with large samples.

It is recommended that various goodness of fit criteria be used together with the overall fit measurement [3,12] in SEM. Some of these measurements constitute indices that generally vary from 0 (inadequate fit) to 1 (perfect fit). The most widely used criteria include: (i) goodness of fit index (GFI): 1 indicates perfect fit; (ii) root mean square error of approximation (RMSEA), in which values less than 0.05 indicate the model's good fit; and (iii) comparative fit index (CFI), in which values greater than 0.9 are expected. Various other indices are available in the statistical packages, including the adjusted goodness of fit index (AGFI), Tucker-Lewis index (TLI) or non-normed fit index (NNFI), standardized root mean square residual (SRMR), and Hoelter's n [12]. More detailed information on goodness of fit criteria can be found in the literature [3,12].

The final stage in the procedure involves examining results that indicate the need for potential modification of the model, in order to maximize the goodness of fit. Modification indices are diagnostic tools that indicate a possible reduction in the chi-square statistic in the overall goodness of fit. These indices can contribute to the evaluation of assumptions adopted in the model. However, the decision to modify it depends mainly on the theoretical implications of such a modification. One of the caveats to the use of these indices is that modified models do not maintain the status of having hypotheses defined a priori, thus implying analyses that are no longer confirmatory, especially if the model is altered substantially [20]. The validity of a model coming from such analyses needs to be confirmed by replication in other data [25].

Structural equation models can be implemented in various statistical packages. These include, among others, AMOS (Analysis of Mo-

ment Structures) [13], an extension module of SPSS (SPSS Inc., Chicago, USA); LISREL (Linear Structural Relationships. Jöreskog K, Sörbom D. Scientific Software International, Lincolnwood, USA), a program developed specifically for the use of SEM, and a pioneer in its application [2,13]; EQS (Multivariate Software Inc., Encino, USA), which incorporates numerous applications related to SEM [2]; MPLUS (Muthén & Muthén, Los Angeles, USA), developed 10 years ago for implementation of SEM; and R (The R Foundation for Statistical Computing, Vienna, Austria; http://www.r-project.org), a statistical program with a specific library for SEM fit [26].

- **Epidemiological study: cognitive development in children under 42 months of age**

Precarious socioeconomic conditions and weak family ties have been identified in the literature as risks for the child's development, thus motivating epidemiological studies on cognitive performance and its relations with the environment and health. Thus, a study was performed to describe the relationship between nutritional status, socioeconomic conditions, quality of home stimulation, and cognitive development in 320 children from 20 to 42 months of age living in the city of Salvador, Bahia State, Brazil [27].

The cognitive development index (CDI) was measured using the Bayley Scales for Infant Development [28]. The scale consists of three complementary subscales (mental, motor, and behavioral), but this study only considered information referring to the mental subscale index, which includes items that evaluate memory, habituation, problem-solving, numerical concepts, generalization, vocalization, language, and social strategies. The characteristics of the home environment and parenting style were measured with Home Observation for Measurement of the Environment (HOME) inventory [29]. Nutritional status was evaluated using the anthropometric scores (z-scores) height/age, weight/age, and birth weight. The procedures for definition of the sample, application of the questionnaires, and description of the measurement scales are described in the literature [27].

The analyses performed previously have used linear regression models [27]. In this study, the data are reanalyzed with structural equation modeling. SEM was used to: (i) define constructs representing characteristics of the environment in which the child lives, parenting style, and the child's nutritional characteristics; and (ii) evaluate the impact of these constructs on the child's cognitive development. The theoretical model for illustrating the use of SEM and explaining the CDI was defined by the researchers in charge of the original study. The path diagram shown in Figure 2 captures the set of causal relations hypothesized by the researchers.

In these analyses, the following variables were used: emotional and verbal responsiveness (*homei*), lack of punishment and restriction (*homeii*), organization of the physical and temporal environment (*homeiii*), availability of appropriate toys and materials for the child's age (*homeiv*), maternal involvement with the child (*homev*), opportunity for variation in the stimulation (*homevi*), height/age anthropometric score (*height/age*), weight/age anthropometric score (*weight/age*), birth weight in kg, sex (1 = female, 0 = male), age in months, and child cognitive development index (CDI). Table 1 shows the description of the CDI and the scores from the HOME Inventory.

As specified in Figure 2, the construct *parenting style* was formed by the scores for the mother's emotional and verbal responsiveness, lack of punishment and restrictions, and maternal involvement with the child. The *physical-environmental* construct was formed by the scores for organization of the physical and temporal environment, availability of toys, and opportunity for variation in stimulation, while the *nutritional* construct was formed by the variables height/age score, weight/age score, and birth weight. The analyses were performed with the R package, version 10.1, and MPlus version 5.21.

Results and discussion

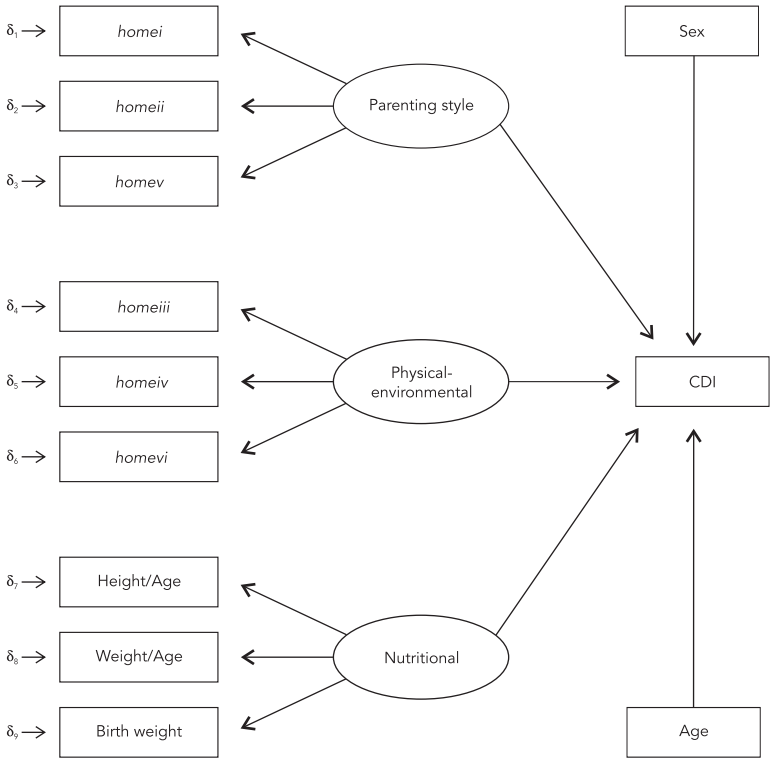
Cognitive development in children under 42 months of age

An assumption associated with use of the maximum likelihood estimation method is the presence of multivariate normality between the observed variables. Preliminary diagnoses for the evaluation of univariate normality did not point to evidence against this assumption. However, evaluation of multivariate normality through the Mardia normality test [30], which evaluates kurtosis and multivariate asymmetry of the distribution, did not support such an assumption. Therefore, the structural equation models fitted to the data considered the use of the maximum likelihood estimator with robust standard errors.

According to the results of the measurement sub-model (Figure 3), one notes that the variable referring to maternal involvement with the child – *homev* – (standardized factor loading = 0.76) exerts a greater contribution to the formation of the construct *parenting style*. The variable refer-

Figure 2

Structural equation model for studying childhood cognitive development.



ring to availability of toys – *homeiv* – (standardized factor loading = 0.79) exerts the greatest contribution to the formation of the construct related to the child's *environment*, while in the *nutritional* construct the variable with the greatest contribution is the weight/age score (standardized factor loading = 0.90). The standardized factor loadings in the measurement sub-model can be interpreted as correlations between the indicators and the corresponding constructs. The score for absence of punishment and restriction – *homeii* – was the only indicator that was not significantly correlated with the construct *parenting style* (standardized factor loading = 0.10; p-value = 0.187). The factor loadings for the other indicators showed moderate to strong magnitude (varying from 0.38 to 0.90).

According to the standardized estimates for the sub-structural model, the *physical-environmental* indicator is the one that most heavily influences the CDI (standardized estimate = 0.50). In addition to the *physical-environmental* factor, the *nutritional* factor (standardized estimate = 0.16) was significantly associated with CDI (Figure 3). The latent variable *parenting style* was not directly associated with CDI (p-value = 0.64), neither were the child's sex and age (respectively, p-values = 0.10 and 0.60).

The chi-square for the overall goodness of fit test was significant (p-value = 0.002), suggesting that the data are not well fitted by the model. However, evaluation using other criteria (RMSEA = 0.049; RMSEA 90%CI = 0.030-0.067); CFI = 0.935; TLI = 0.906; SRMR = 0.041) indicates that the model fits the data reasonably well. The 0.067 value at the upper limit of the 90% confidence interval for RMSEA indicates that the model can be improved if modifications are made. The modification indices suggest the inclusion of a relation between the *homei* and *homev* indicators and the *physical-environmental* construct.

The proposed modifications were performed, but resulted in problems of identifiability or inadequate values for the models' goodness of fit

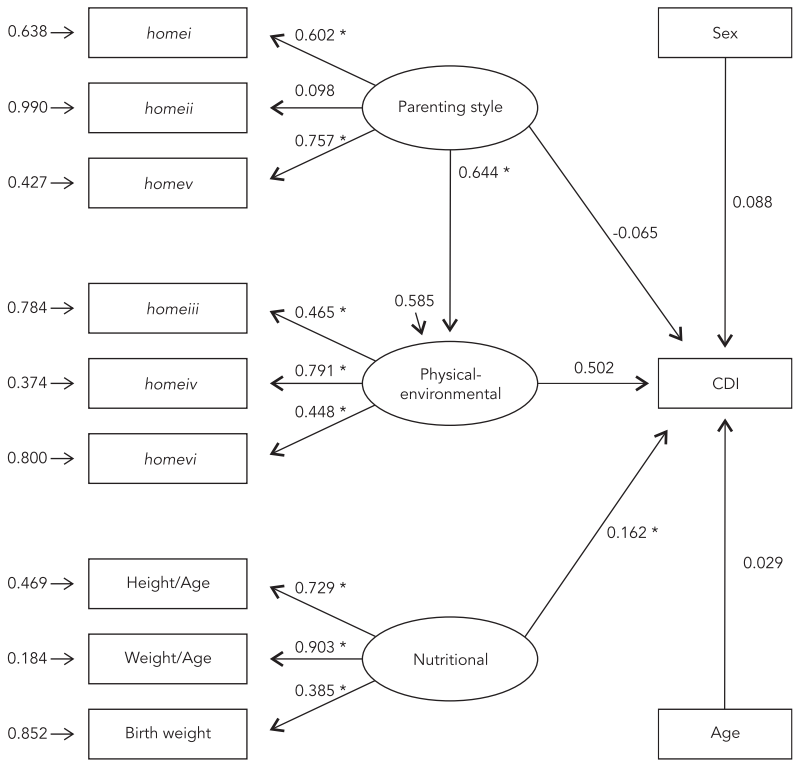
Table 1

Description of the cognitive development index and the scores from the HOME inventory used in the example of application.

Name	Description	Range	Interpretation
CDI	Cognitive development index	70-120	Higher levels indicate higher level of development
Homei	Emotional and verbal responsiveness score	2-11	Higher values indicate higher level of responsiveness
Homeii	Score for absence of punishment and restriction	0-8	Higher values indicate less punishment and restriction
Homeiii	Score for organization of physical and temporal environment	0-6	Higher values indicate higher level of organization in the child's routine
Homeiv	Score for availability of adequate toys and materials for the child's age	0-9	Higher values indicate greater availability of adequate toys and materials for the child's age
Homev	Maternal involvement score	0-6	Higher values indicate greater maternal involvement with child
Homevi	Score on opportunity for variation in stimulation	0-5	Higher values indicate more opportunities for variation in the child's daily stimulation

Figure 3

Estimates from the structural equation model for studying childhood cognitive development.



* p-valor < 0.05.

evaluation indices. However, removal of the variables sex and age from the sub-structural model resulted in an appropriate model, displaying the following goodness of fit statistics: chi-square = 37.6 with 31 degrees of freedom (p-value = 0.19); RMSEA = 0.026 (90%CI: 0.000-0.053); CFI = 0.986; TLI = 0.980; and SRMR = 0.036.

Figure 4 shows the estimates for the final model. No substantial changes were verified in the estimates presented in the previous model. The R^2 statistic associated with the model varied from 0.01 (for the variable *homeii*) to 0.76 (for the weight/age score), describing the amount of variance explained by the corresponding construct. Except for the variable *homeii*, the R^2 values indicate that the model is capturing the variances of the observed variables reasonably well. In addition to the positive relationship observed between the *nutritional* and *physical-environmental* constructs with CDI, the total estimated effect for the construct *parenting style* in the CDI was also positive and significant (standardized estimate = 0.259; standard error = 0.072; p-value = 0.000). The adjusted model allowed the breakdown of the total effect of *parenting style* into direct effect (standardized estimate = -0.065; standard error = 0.141; p-value = 0.650) and indirect effect (standardized estimate = 0.323; standard error = 0.103; p-value = 0.002), suggesting an important mediating role by the *physical-environmental* construct in this relationship.

Considering the results obtained in the analyses, the positive impact of the quality of psychosocial stimulation existing in the home environment on cognitive performance shows that part of the effect of stimulation on CDI is due to the *parenting style* of interaction with the child and to the *physical-environmental* characteristics of the family context. This finding is consistent with other studies in the area [27,31,32]. However, the current study did not consider broader social characteristics like maternal occupation and schooling, which are known to influence the acquisition of cognitive skills in childhood [27,31].

Applying a different analytical model from that used previously for the same data, one identifies more accurately the specific components of home stimulation on cognitive development in early childhood. *Parenting style*, a fundamental element in the child's emotional development and personality [33,34], showed a relatively smaller influence on CDI than the availability of toys and variation in the home routine on CDI. However, the use of SEM favored the understanding that *parenting style* can exert an influence on cognition through organization of the child's environment, variation in daily stimulation, and availability of adequate toys. Mothers that are more responsive and involved in their children's development may be the ones that offer greater opportunities for interaction and complexification of the child's environment, thereby favoring childhood cognitive development.

Another gain with the use of SEM in this epidemiological study was the identification of two theoretical constructs (parenting style and quality of the child's physical home setting) based on the subscales in the HOME inventory. This procedure allowed a theoretical and analytical adjustment of the established relations and greater precision in the estimates, with evaluation of each construct's contribution to explanation of the CDI. In addition, use of the *nutritional* construct in SEM was able to identify an association between nutritional aspects and cognitive development in early childhood, which was not observed in the analysis with more traditional methods that consider each variable separately [27]. SEM demonstrated its potential for the analysis of epidemiological data, especially when the object requires a more complete understanding of the interrelations between different target factors, as in the case of the current study.

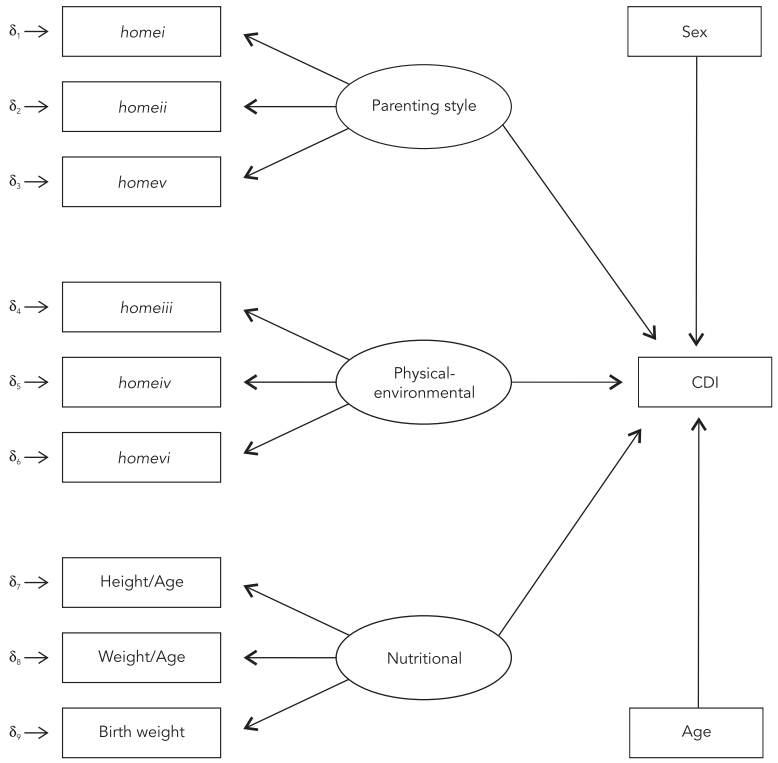
Final remarks

The aim of this study was to provide a summary of the basic principles and concepts in structural equation modeling (SEM) and to present an empirical example of this type of analysis. Numerous applications in the field of Epidemiology can benefit from this methodology, including examination of the number of dimensions that comprise the target phenomena and the validation and evaluation of the reliability of measurement scales. A specific area of application is in behavioral genetics studies [16], which evaluate the contribution of genetic factors to the variation in observed phenotypes. The assumption that measurements are performed without error may not be realistic in many of these applications. One of the strongest justifications for introducing SEM is modeling of the causal relationship, since these models constitute powerful tools for expressing causal theories, allowing their verification through the model's mathematical fit and determination of the extent to which the observed data provide evidence that supports the theory. In addition, simulation studies [35] indicate that SEM is superior for studying mediation, when compared to traditional regression models.

This study approached SEM involving continuous endogenous variables, with a consolidated statistical theory that is already widely used in various fields of knowledge. However, advances

Figure 4

Estimates from the reduced structural equation model for studying childhood cognitive development.



* p-valor < 0.05.

in SEM have allowed the inclusion of new estimation techniques to deal with non-normal distributions. Due to the work of Muthén [36,37] and others, it is possible to estimate the complex SEM parameters when the data are non-normal, including the mixture of dichotomous, ordinal categorical, and continuous variables. There is also a SEM literature for binary responses or those involving contexts in which the multiple normality assumption does not apply [37,38]. For the case of binary responses, Kupek [38] proposed the use of Yule transformation to obtain the odds ratio to describe and interpret the interrelations between variables in SEM.

Another expanding area is the use of SEM for situations involving complex data structures with correlation between multiple observations of the same individual or between individuals belonging to the same group or cluster. The application of SEM to growth curves or using the multilevel modeling approach has attracted the interest of various researchers [39,40,41]. The use of this methodology helps overcome the limitations of traditional methods, allowing progress in epidemiological knowledge. Such analysis of a theoretical model that involves a more complex causal network, including difficult-to-measure variables, is advantageous for minimizing the residual confounding related to the principal association, especially in observational studies, where there is a limitation due to the impossibility of completely controlling the confounding variables.

Although SEM is not used frequently for analyzing epidemiological data, the discussion on this topic has existed for some time, especially in the debates on causal modeling and the role of Statistics in causal inference [42]. Importantly, however, like any procedure in data analysis, this methodology is also subject to misspecifications, and can be considered concurrently with other procedures to allow a more robust evaluation of the target interrelations.

Resumo

A modelagem de equações estruturais (MEE) é uma ferramenta estatística importante para avaliar relações complexas em várias áreas do conhecimento. Em Epidemiologia sua divulgação e uso são limitados. Este artigo apresenta princípios e conceitos básicos da MEE, com exemplo de aplicação na análise de dados epidemiológicos. A análise de dados é realizada em estudo que investiga determinantes do desenvolvimento cognitivo infantil, sendo definidos construtos relacionados à organização do ambiente da criança, ao seu status *de saúde, e às práticas e estilo de vida dos pais. O impacto positivo da qualidade de estimulação psicossocial do ambiente doméstico sobre o índice de desempenho cognitivo (IDC) esclarece que parte do efeito da estimulação sobre o IDC deve-se ao estilo parental de interação com a criança e às características físico-ambientais do contexto familiar. As potencialidades do uso da MEE em Epidemiologia são apresentadas, enfatizando-se a definição do modelo teórico e seu uso para aprofundamento de questões epidemiológicas sob nova perspectiva.*

Modelos Matemáticos; Análise Fatorial; Causalidade

Contributors

L. D. A. F. Amorim and L. T. L. P. Moraes participated in the elaboration, implementation, and interpretation of the statistical analysis. They contributed to the final drafting and revision of the article. R. L Fiaccone participated in the final draft and revision of the article. C. A. S. T. Santos and S. O. Barbosa participated in the elaboration, implementation, and interpretation of the statistical analysis and contributed to the final revision of the article. T. N. Santos and N. F. Oliveira contributed to the discussion of the statistical methodology, literature search, and final revision of the article. D. N. Santos, L. M. Santos, S. M. A. Matos, and M. L. Barreto collaborated in the elaboration of the data collection instrument and the data survey in the epidemiological study, definition of the theoretical conceptual models for the epidemiological studies, discussion, and interpretation of the results. They also contributed to the final revision of the article.

References

1. Kaplan D. Structural equation modeling: foundations and extensions. Thousand Oaks: Sage Publications; 2000. (Advanced Quantitative Techniques in the Social Sciences Series, 10).
2. Raykov T, Marcoulides GA. A first course in structural equation modeling. 2nd Ed. New York: Psychology Press; 2006.
3. Strumacker E, Lomax RG. A beginner's guide to structural equation modeling. 2nd Ed. New Jersey: Lawrence Erlbaum Associates; 2004.
4. Bollen K. Total, direct and indirect effects in structural equation models. Sociol Methodol 1987; 17:37-69.
5. Ditlevsen S, Christensen U, Lynch J, Damsgaard T, Keiding N. The mediation proportion: a structural equation approach for estimating the proportion of exposure effect on outcome explained by an intermediate variable. Epidemiology 2005; 16:114-20.
6. Shar S, Novak S, Stapleton LM. Evaluation and comparison of models of metabolic syndrome using confirmatory factor analysis. Eur J Epidemiol 2006; 21:343-9
7. Olson K, Hayduk L, Cree M, Cui Y, Quan H, Hanson J, et al. The changing causal foundations of cancer-related symptom clustering during the final month of palliative care: a longitudinal study. BMC Med Res Methodol 2008; 8:36.
8. Silva AAM, Mehta Z, O'Callaghan FJK. Duration of breast feeding and cognitive function: Population based cohort study. Eur J Epidemiol 2006; 21: 435-41.
9. Dahly DL, Adair LS, Bollen KA. A structural equation model of the developmental origins of blood pressure. Int J Epidemiol 2009; 38:538-48.
10. Bollen K. Structural equations with latent variables. New York: Wiley; 1989.
11. Farias SA, Santos RC. Modelagem de equações estruturais e satisfação do consumidor: uma investigação teórica e prática. Revista de Administração Contemporânea 2000; 4:107-32.
12. Hair JF, Anderson RE, Tathan RL, Black WC. Análise multivariada de dados. 5ª Ed. Porto Alegre: Bookman; 2005.
13. Kline RB. Principles and practice of structural equation modeling. New York: The Gilford Press; 2005.
14. Codes ALM. Modelagem de equações estruturais: um método para a análise de fenômenos complexos. Caderno CRH 2005; 18:471-84.
15. Chavance M, Escolano S, Romon M, Basdevant A, Lauzon-Guillain B, Charles MA. Latent variables and structural equation models for longitudinal relationships: an illustration in nutritional epidemiology. BMC Med Res Methodol 2010; 10:37-63.

16. Fergusson DM. Annotation: structural equation models in development research. J Child Psychol Psychiat 1997; 38:877-86.
17. Jorëskog KG. Testing a simple structure hypothesis in factor analysis. Psychometrika 1966; 32:165-78.
18. MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariances structural modeling. Psychol Methods 1996; 1:130-49.
19. Mueller RO. Basic principles of structural equation modeling. New York: Springer-Verlag; 1996.
20. Ullman JB. Structural equation modeling: reviewing the basics and moving forward. Journal of Personality Assessment 2006; 87:35-50.
21. Grace JB, Bollen KA. Interpreting the results from multiple regression and structural equation models. Bulletin of the Ecological Society of America 2005; 40:283-95.
22. MacKinnon DP. Introduction to statistical mediation analysis. New York: Lawrence Erlbaum Associates; 2008
23. MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. Psychol Methods 2002; 7:83-104.
24. Jorëskog KG. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 1969; 34:183-201.
25. MacCallum R. Specification searches in covariance structure modeling. Psychol Bul 1986; 100:107-20.
26. Fox J. Structural equation modeling with the SEM package in R. Struct Equ Modeling 2006; 13: 465-86.
27. Santos LM, Santos DN, Bastos ACS, Assis AMO, Prado MS, Barreto ML. Determinants of early cognitive development: hierarchical analysis of a longitudinal study. Cad Saúde Pública 2008; 24: 427-37.
28. Bayley N. The Bayley scales of infant development. 2nd Ed. New York: Psychological Corporation; 1993.
29. Caldwell B, Bradley RH. Home Observation for Measurement of the Environment (HOME): revised edition. Little Rock: University of Arkansas Press; 1984.
30. Mardia KV. Measures of multivariate skewness and kurtosis with applications. Biometrika 1970; 57:519-30.
31. Andrade SA, Santos DN, Bastos AC, Pedromônico MRM, Almeida-Filho N, Barreto ML. Ambiente familiar e desenvolvimento cognitivo infantil: uma abordagem epidemiológica. Rev Saúde Pública 2005; 39:606-11.
32. van Bakel H, Riksen-Walraven M. Parenting and development of one-year-olds: links with parental, contextual and child characteristics. Child Development 2002; 73:256-73.
33. Vitolo YLC, Fleitlich-Bilyk B, Goodman R, Bordin IAS. Crenças e altitudes educativas dos pais e problemas de saúde mental em escolares. Rev Saúde Pública 2005; 39:16-24.
34. Ferreira MCT, Marturano EM. Ambiente familiar e os problemas do comportamento apresentados por crianças com baixo desempenho escolar. Psicol Reflex Crit 2002; 15:35-44.
35. Iacobucci D, Saldanha N, Deng X. A meditation on mediation: evidence that structural equations models perform better than regressions. J Couns Psychol 2007; 17:139-53.
36. Muthén BO. A general structural equation model with dichotomous, ordered categorical and continuous latent indicators, Psychometrika 1984; 49:115-32.
37. Muthén BO. Beyond SEM: general latent variable modeling. Behaviormetrika 2002; 29:81-117.
38. Kupek E. Beyond logistic regression: structural equation modelling for binary variables and its application to investigating unobserved confounders. BMC Med Res Methodol 2006; 6:13.
39. Bauer DJ. Estimating multilevel linear models as structural equation models. J Educ Behav Stat 2003; 28:135-67.
40. Singer J, Willett JB. Applied longitudinal data analysis: modeling change and event occurrence. London: Oxford University Press; 2003.
41. Rabe-Hesketh S, Skrondal A, Pickles A. Generalized multilevel structural equation modeling. Psychometrika 2004; 69:167-90.
42. De Stavola BL, Nitsch D, Silva IS, McCormack V, Hardy R, Mann V, et al. Statistical issues in life course epidemiology. Am J Epidemol 2006; 163: 84-96.

Submitted on 08/Nov/2009
Final version resubmitted on 27/May/2010
Approved on 30/Jun/2010